CONFIGURATIONAL vs. TEMPORAL COHERENCE IN AUDIOVISUAL SPEECH PERCEPTION

M.-A. Cathiard, M.T. Lallouache°, T. Mohamadi° and C. Abry°*
**Laboratoire de Psychologie Expérimentale, Université Mendès-France, BP 47 F-38 040 Grenoble Cedex 9 /°ICP URA CNRS n° 368 INPG/ENSERG Université Stendhal BP 25*

ABSTRACT

This contribution evaluates the natural *coherence* of the building up of audiovisual information in the flow of speech that is provided on the face. Our conclusion is that the bimodal coherence of speech is above all configurational, and not essentially temporal, as it is the case in another intersensory system, such as auditory-visual localization in the ventriloquist effect [1, 2].

1. INTRODUCTION

The aim of this contribution is to evaluate the natural coherence of the building up of visual *and* audio information in the flow of speech, using a gating and desynchronization procedure.

It has been previously shown that, when optic information is naturally in advance on acoustics – as in the case of a visible gesture like rounding –, featural/gestural anticipation can be identified only by eye several tens of milliseconds before any perceivable sound [3].

This bimodal information being in a sense naturally "desynchronized", an obvious experimental manipulation was to reduce the delay of audition on vision, i.e. to "resynchronize" the audio signal in order to test the boundaries of such a bimodal temporal organization.

2. STIMULI

We used [i#y] and [i#i] (control) vocalic transitions embedded in a carrier sentence of the type: "T'as dit : UHU use?" ("Did you say : UHU ["Indian" name] wear out?"). Indian names were used in order to maximize pure vowel-to-vowel modulation of the output of the vocal-tract, without intervening consonantal gestures. A French male talker was filmed at 25 frames/second. Stimuli were chosen with a 160 ms acoustic pause between vowels. For each digitized frame, articulatory parameters were automatically extracted by image processing [4].

For [i#y] transitions, upper lip protrusion P1 starts at the end of [i] (Fig. 1), together with the constriction, i.e. lip area S begins to decrease. The domain we determined, for these two main components of the rounding gesture, stretches from 4 images before the pause, in order to allow a sufficient range in desynchronization, to 1 image after, when *both* components have reached their maximum.

3. VISUAL TEST

The visual information was explored by gating (40 ms steps): 10 gates of 1000 ms duration allowed to display in all cases the beginning of the carrier sentence "T'as dit...". This visual [i]/[y] identification test (with 10 repetitions for each gate in random order) was performed by 10 naive French subjects, with no deficit in vision and audition. The visual boundary in the [i#y] transition was measured on the mean curve of all subjects using Probit analysis.

This boundary takes place 140 ms before the acoustic onset of [y], and less than 40 ms (a one image step) are enough to switch from 80% [i] to 80 % [y] (Fig. 1). Anticipation is earlier (140 *vs.* 100) and category switching is steeper (40 *vs.* 80) than the one we obtained with another 160 ms paused signal, whose articulatory profile was actually slower, especially in the constriction building up [5]. Anyway, this 140 ms anticipation remains shorter than the maximum case we ever evidenced, in fact within a very long 460 ms pause: 210 ms [5].

Thus we confirm our previous findings on the natural advance of the eye relative to the ear.

4. AUDIO TEST

So what about the building up of audio information ?

An audio [i]/[y] identification test, including the beginning of the stimuli "T'as dit...", stopped 2, 6 and 10 ms after the onset of [#y] and [#i], was performed by the same 10 subjects (10 repetitions by gate). Such a range from 2 to 10 ms has proven to be sufficient to scan properly the building up of featural acoustic information. Mean identification scores were: 58%, 95%, 99% for [y] (Fig. 2); and 100%, 99% et 99% for [i]. This supports the claim that often only one pulse (8 ms in the case of the vowel onsets of our talker) is sufficient to fairly identify the vowel (for French, see [6]).

The building up of visual information (40-80 ms) is thus slower than the audio one (10 ms). But this is fairly compensated by the visible anticipation on the sound, which is naturally displayed in speech (up to 200 ms) due to such a pervasive phenomenon as coarticulation.

5. AUDIOVISUAL TEST

But what are the audio/visual boundaries of this bimodal coherence ?

The same visual stimuli were presented with the sound in synchrony or in advance. For each gate, for which we measured the time course of visual information, we tested the building up of the audio using the 3 steps previously determined (2, 6 et 10 ms), in order to obtain a desynchronization range from 0 to -360 ms, by 40 ms steps. The same 10 subjects where tested on the 10 steps for the 3 vowel onset durations.

Individual curves obtained for each acoustic duration show clearly different patterns. Since averaging was not representative, we grouped them according to similarity of their response profiles. Individual curves are either clearly S-shaped, or they show a first phase, before the visual [i#y] boundary, which is less regular and/or close to chance level (Figs. 3a-f). On the base of the scores for 10 ms vowel onsets (corresponding to high audio performances), we obtained two groups of 5 subjects. The first group (Figs. 3a-c) has, in the phase before the visual boundary, identification [y] scores below 20%; the second group having scores above 20% (Figs. 3d-f).

If we first consider scores in the synchronous condition (plotted on gate n°10) for both groups, we see that, independently of vowel onset duration, individual [y] scores are generally at or above 90% (with one exception). Mean scores for 2, 6 et 10 ms durations are respectively: 96%, 98% et 99%. Comparing audiovisual results obtained in the worst condition, 2 ms (96%), with the audio alone condition (58%), vision benefit reveals largely sufficient to disambiguate a poor audio signal. We thus rejoin results in a more classical condition, namely speech in noise (for French, see [7]).

When desynchronization occurs, for these 2 ms vowels, we see that rounding information – an anticipating one in the original – can bring to them a visual benefit up to -160 ms. One must recall that for this value, i.e. up to image n°6 (see Fig. 1), visual information alone reached 85% [y] responses, whereas just 40 ms before it scored 12%. In other terms, we were able to test step by step what phase of the anticipatory gesture could enhance ambiguous audio information. It comprises in fact all the phase "sheltered" by the gesture: *after the visual boundary.*

Let's consider now desynchronization effect beyond this visual boundary, i. e. for the phase corresponding articulatorily to an [i]. For Group 1, we see that the duration of vowel [y] onsets – comparatively to identification scores in the audio condition – does not seem to influence subjects' behaviour. In fact, what is properly characteristic of this group is its high sensitivity to visible articulatory information. The curves we obtain for the three conditions display a clear S-shape, which looks strongly like the ones (mean and individual) obtained for vision alone: the identification boundary is located, for the three audiovisual conditions in the vicinity of the visual boundary. This similarity of the curves in the visual and audiovisual conditions indicates that, when desynchronization delivers images in advance of the sound – in this case an articulatory information specific of an [i] (in a desynchronization ranging from -200 to -360 ms, for this transition) –, then subjects identify the oncoming of an [i] vowel, in spite of the fact that they receive an audio information largely sufficient to recognize an [y]. Things are going on as if in the case of conflicting information – [i] being visible et [y] audible –, visible information was guiding perception.

Subjects from Group 2 are sensitive also to conflicting information. Whereas audio is clearly identified by them as [y] – at least for 6 and 10 ms durations –, a visible [i] pushes their scores towards chance level.

To summarize: (i) desynchronization has the largest effect only when the visual boundary is crossed; (ii) beyond this boundary, no subject is insensitive to visual information, i.e. clearly no subject displays a steady 100% [y] along all desynchronization values. Moreover in (ii) the proportion of those who answer [i] for audio [y] is very close to the results found in a rounding judgment task for the same conflicting French vowels [8]. However, up to the present experiment, no such "McGurk effect" had been successfully obtained by a desynchronization procedure for vowels [9].

6. CONCLUSION

The natural delay of audio, relative to the visual signal, in speech coarticulatory anticipation, can be reduced without affecting intelligibility, as long as the configurational visual cues are in accordance with the sound. This hypothesis of a primacy of configurational over temporal coherence could be used to explain other results on desynchronization (reviewed in [10, 3, 11]) for detection tasks [12] as well as for intelligibility ones [13, 14].

Acknowledgements: This work was supported in part by a grant in Cognitive Science from the French Ministry of Research and Technology, a France Telecom-Université Stendhal contract n°927B032 and an Esprit Basic Research project n°6975 *Speech Maps*.

REFERENCES

[1] Radeau, M. (1994). Auditory-visual spatial interaction and modularity. *Current Psychology of Cognition, 13(1)*, 3-51.

[2] Abry, C., Cathiard, M.-A., Robert-Ribès, J., & Schwartz, J.-L. (1994). The coherence of speech in audio-visual integration. *Current Psychology of Cognition, 13(1)*, 52-59.

[3] Cathiard, M.-A. (1994). *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole.* Thèse de Psychologie Cognitive, Université Grenoble 2.

[4] Lallouache, M.-T. (1991). *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres.* Thèse de l'ENSERG, Spécialité : Signal Image Parole, Grenoble.

[5] Cathiard, M.-A., & Lallouache, T. (1992). L'apport de la cinématique dans la perception visuelle de l'anticipation et de la rétention labiales. *Actes des 19èmes Journées d'Études sur la Parole*, Bruxelles, 19-22 Mai, 25-30.

[6] Serniclaes, W., & Wajskop, M. (1972). L'identification vocalique en fonction de la fréquence fondamentale et de la durée de présentation. *Revue de Phonétique appliquée, 22*, 39-50.

[7] Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37*, 1195-1203.

[8] Lisker, L. & Rossi, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech, 35(4)*, 391-417.

[9] Massaro, D.W., & Cohen, M.M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*, 127-134.

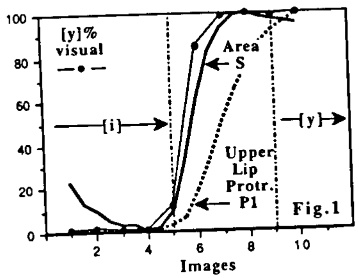
[10] Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In V. Bruce, A. Cowey, A.W. Ellis & D.I. Perrett (Eds.), *Processing the facial image* (Proceedings of a Royal Society Discussion Meeting, 9-10 July), Clarendon Press, Oxford, pp. 71-78.

[11] Cathiard, M.-A., & Tiberghien, G. (1994). Le visage de la parole : une cohérence bimodale temporelle ou configurationnelle? *Psychologie Française, 39 (4)*, 357-374.

[12] McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77(2)*, 678-685.

[13] Smeele, P.M.T., & Sittig, A.C. (1991). The contribution of vision to speech perception. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genova, Italy, 24-26 September, vol. 3, 1495-1497.

[14] Smeele, P.M.T. (1994). *Perceiving speech : Integrating auditory and visual speech.* Doctoral dissertation, Delft University.



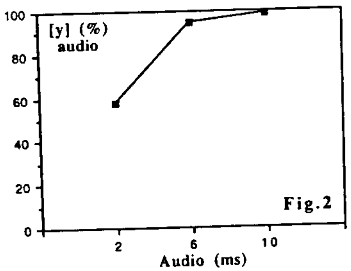
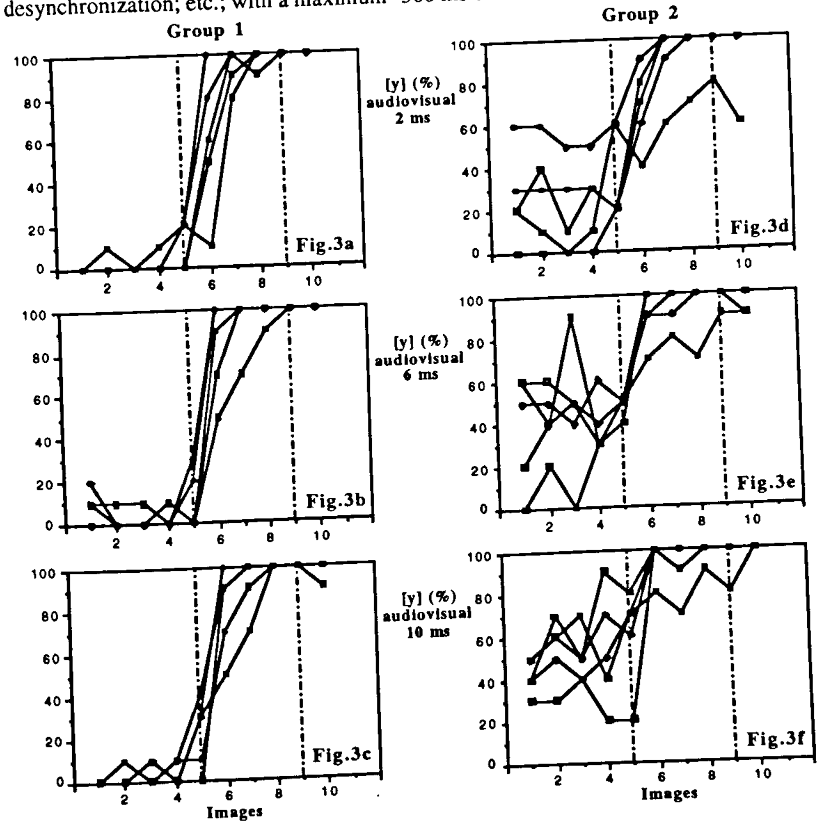


Fig.1.– Mean visual identification results for [tadi#y] superimposed on normalized articulatory parameters -S and +P1 (scores are plotted on the last image of the gate). **Fig.2.**– Mean auditory scores for [tadi#y] with short [y] onsets. **Figs.3a-f.**– Individual audiovisual scores for group 1 (left) and group 2 (right). N°10: scores in the synchronized condition with (from top to bottom) 2, 6 and 10 ms [y] onsets after image n°9. For scores on n°9 the sound is aligned on n° 8, i.e. with a -40ms desynchronization; etc.; with a maximum -360 ms on n°1.


Group 1
Group 2
[y] (%)
audiovisual
2 ms
[y] (%)
audiovisual
6 ms
[y] (%)
audiovisual
10 ms
Fig.3a
Fig.3b
Fig.3c
Fig.3d
Fig.3e
Fig.3f
Images